Batanero, C. (Ed.), Training Researchers in the Use of Statistics, 119-128.
© 2001 International Association for Statistical Education and International Statistical Institute.
Printed in Granada, Spain.

SANDRA MCDONALD

PRACTICAL AND EDUCATIONAL PROBLEMS IN SHARING OFFICIAL MICRO DATA WITH RESEARCHERS

Many commentators have noted the need for reform in statistical education. They tend to focus on the analytical techniques that are critical to understanding and producing good quality statistical outputs. This paper adds to these discussions and looks at some of the main analytical issues that transpire from researchers accessing the particular form of statistical data sets in a national statistical office. However the conclusions are much more widely applicable to other data sets. It also considers the more practical, but very important skills and knowledge, applicable to all types of data, that a researcher needs, such as fashioning the data set into a format that is most useful to them, and ensuring they obtain access to data that will allow them to fulfil their research objectives.

1. INTRODUCTION

In this paper I focus on researchers' experiences with using micro data collected for the purpose of generating official statistics, basing that on the experiences of Statistics New Zealand which has recently provided researchers with access to micro data. It provides insights into the specific pitfalls and problems that can be experienced both by the researcher and a national statistical office where official statistical data sets are being used. There are recommendations that both parties can implement to facilitate smoother working relationships.

However the ideas and conclusions do not have to be limited to just the use of official statistical data sets. Many of the experiences can be applied to other data sets, especially where researchers are using data that are collected by somebody else. These are increasingly being made available for wider use by institutions such as universities and data archives like the UK Data Archive. While the data appears to be ready to use, researchers often needs to invest almost as much time and energy in familiarising themselves with the data as they would if they were collecting their own.

2. STATISTICS NEW ZEALAND'S ENVIRONMENT

The collection of data required to produce official statistics is expensive and a significant burden on individual and business respondents. To ensure society gets sufficient return for the cost it is important to make the best use of the data.

In common with national statistics offices (NSO's) around the world, Statistics New Zealand (SNZ), for reasons such as resource constraints, competing priorities and arguably tradition, does not undertake as much in-depth analytical work as it would like.

Besides, it is not always desirable for all analyses of official statistical data collections to be carried out by an NSO. As Biggieri and Zuliani (1999) note NSO's have staff with good quantitative skills but who aren't always so in touch with the sorts of policy issues or decisions that their data sets could be applied to.

To facilitate contributions by other researchers an NSO traditionally makes a wide range of aggregated output available. Increasingly though it is becoming more common for an agency to make unit record data more accessible to researchers through micro data sets. Benefits of access to micro data for researchers include the ability to recode subset and sort the data, derive new variables, and deal appropriately with outliers, but most importantly micro data is essential for using multivariate techniques.

SNZ's legislation does not allow it to publicly release confidentialised micro data but it does permit provision of access to unidentified micro data under certain conditions. This has resulted in the particular form of statistical micro data access available in New Zealand.

The Data Laboratory was established in 1997 to provide a transparent process for managing access to micro data to approved researchers. The objective is to assist SNZ to increase the value obtained from its data sets. For the three years it has been operating the Data Laboratory has successfully resulted in a significant increase in the amount of research undertaken. Prior to the Data Laboratory, SNZ micro data was used infrequently, once or twice a year at most. More recently access has been provided to micro data for up to 15 projects each year.

Overall the Data Laboratory has greatly increased the opportunities for the government sector and the research sector to collaborate for the benefit of public research. It has, however, also highlighted some aspects that researchers using official statistical data sets have found difficult.

These are discussed in detail in the following sections and lead to the conclusions and recommendations for training of social science researchers undertaking research using the statistical data sets collected by other agencies. In its turn, SNZ has identified areas where it can instigate improvements to make the work of researchers' a lot easier.

Table 1: Use of the Data Laboratory (since July 1997)

Types of researchers	Number of projects	Data used
Government departments (nine of the projects were contracted to academic researchers)	23	Population census (4) Household economic survey (4) Household labour force survey (4) Other household surveys (7) Business data (5)
Academic researchers (including post-graduate students)	11	Population census (4) Household economic survey (3) Other household surveys (4)
Research institutes/ independent researchers	2	Population census (1) Household economic survey (1)
Local government	1	Population census (1)

3. UNDERSTANDING ISSUES THAT ARE CRITICAL TO A NATIONAL STATISTICS OFFICE

a. Respondent trust

While the Statistics Act provides SNZ with the power to compulsorily collect data, it

is preferable for the department for respondents to willingly provide information without the need to resort to the compulsory provisions. To maintain high response rates, and therefore provide data collections that are widely regarded as providing high quality outputs, a NSO needs to be trusted by the public, and in particular by its respondents. Any diminution of that trust would quickly affect the quality of the data it collects.

SNZ achieves good response rates. Apart from the obvious benefits to data quality it reduces the need for substantial work that would be required to compensate for low response rates, so reduces potential costs.

It is important, therefore, for researchers to recognise the constraints on policies and operational activities where statistical agencies rely heavily on public trust and good will, and how quickly that trust can be undermined if an individual respondent's privacy is breached. There are a number of occasions where, through loss of trust, a national agency's reputation has been impacted overnight and has only been retrieved over a long period of time and with a great deal of money.

A NSO offering access to micro data needs to be clear about its position on data access within the context of the wider research environment. SNZ's approach is to provide access to micro data in a way that ensures there is little risk of disclosure at an individual level. Where there is the potential for conflict between protecting respondents' information and a research objective SNZ will act conservatively and choose the protection of the respondent. However this may mean that there are constraints on what can be made available to a researcher that have more to do with public perception than with any specific concerns about confidentiality.

Researchers who are exclusively focused on their research objectives may not understand or accept the need for such a stance. While they are unlikely to want the NSO to act in any way that would be detrimental to the quality of what it collects, researchers may not always agree that their proposal will have an adverse impact. A researcher, therefore, needs to be aware of where their research fits within the wider context of the NSO's operations.

b. Reducing disclosure risk in a unit record data set

When a researcher requests access to micro data, SNZ asks them to carefully consider the data they require and to justify the inclusion of variables in the subset of the data they get access to. By providing access only to the data researchers need for their research and not the full data set it assists with minimising disclosure risk. Researchers will not get approval for access to micro data for a project that appears to be a "fishing expedition", where they want all available variables and do not have a well-defined research problem.

Glencross and Mji (2001) and Bishop and Talbot (2001) both identify the importance of this stage. Glencross and Mji describe it as formulating the research problem, involving two key tasks of identifying the 'what' and the 'why'. Where a research project involves the use of data that has been collected by some else, whether by an NSO or some other agency, as opposed to collecting their own there are issues that a researcher needs to consider.

To obtain approval for access to SNZ's micro data researchers need to be very clear about the outcomes they are trying to achieve and to show that the data they are requesting will assist them to attain those outcomes. They should ask appropriate questions to find out what they need to know about the data. This can be a "chicken and egg" situation if they don't have the data in front of them.

The agency holding the data set must make sufficient documentation about the data set available and be prepared to spend some time answering queries about the data collection and how the data set is structured.

Another area that involves negotiation with the researcher is the process of modifying a data set to reduce the risk of disclosure. Because SNZ cannot provide public access to micro data we develop a data set specific to the requirements of each researcher. While we have some standard approaches to reducing disclosure, such as limiting regional detail or providing age rather than birth date, other steps will depend on what data the researcher has requested and also the sensitivity of the data itself. This process is undertaken in consultation with the user to reduce potential adverse impacts on their research as much as possible.

c. Confidentialising output

An issue to emerge from SNZ's experiences with researchers' use of micro data is their different perspective on confidentiality. At the beginning of any project involving the use of micro data we have found it important to ensure that the researcher understands the importance placed on confidentiality to preserve respondent trust.

Researchers often assume that if the data set does not contain names and addresses there is no need for any further confidentiality protection. They are not immediately aware of how easily disclosures can occur in output that is not adequately confidentialised (for example, when a table contains a cell with a single entry or how an individual's information can be disclosed by decomposing data across several independently produced tables).

SNZ realises that its own staff do not find confidentiality an easy concept to comprehend so it is not surprising that researchers are not knowledgeable about it. To assist researchers SNZ has prepared documentation that explains the theory behind the need for the rules that are in place and describes the different types of confidentiality techniques that are needed for different types of data sets, for example censuses as opposed to samples and household as opposed to business data.

When they are using SNZ micro data, researchers need to become familiar with how the department applies confidentiality rules to its data, as they are required to apply the same techniques to their output. This is made easier for the researcher as we provide them with programs and macros developed in-house to ensure that what they do is consistent with our practice. It also reduces the workload of SNZ staff checking confidentiality of output if researchers use standard output protection techniques.

While I am not proposing that researchers should be trained in specific techniques they do need to understand the general issue of confidentiality and to be in a position to appreciate the critical importance of confidentiality to a NSO.

Interestingly the use of micro data by external researchers has highlighted possible inconsistencies with SNZ's output practices. When output made available by SNZ was limited to aggregated tables it was difficult for users to challenge the rules that were applied.

Users were not always in a position to determine whether they were sensible and also whether they resulted in data being suppressed unnecessarily. Through needing to explain, and defend, the rules to researchers who quite rightly question aspects that seem inappropriate, SNZ is much more aware of the need to be clear and consistent in its own practices.

4. COMPLEXITY OF OFFICIAL STATISTICAL DATA SETS

The stage of the research process that involves researchers in organising the collection of data – the D (Do) part of Bishop and Talbot's (2001) PPDSA cycle - must, in cases where they are using existing data sets, be replaced with interrogation of collection documentation and discussions with the collection agent, which for official statistics is the NSO.

In SNZ's experience, researchers find official statistical data sets much more complex than they expect. Common feedback is that there is a steep learning curve to becoming sufficiently familiar with a data collection to be in a position to apply appropriate analytical techniques and to obtain meaningful results. Researchers using official statistical micro data also need to accept that, in the short-term, they are likely to have little influence over data collections. However, as the NSO is exposed to the policy issues and decisions, their data sets are used for useful changes and may feed back into decisions about collections in the long-term.

As Glencross and Mji (2001) note for the education of social science researchers to be effective it needs to relate to the context of the research. Unless data sets that are relevant to their course are available researchers' training is can be restricted by "reliance on a small number of rather tired old data sets which have been used extensively on many courses" (Chant & Lievesley, 1997). Where students collect their own data the data set is likely to be very small and simple. It is therefore unlikely to have many of the problems inherent in real data set, such as a large official statistical data set, which will almost certainly contain imputed and edited data. A researcher may be unaware of the time and effort they will need to invest to produce a data set in a format that they is suitable for their research purposes.

It would be useful for a researcher, particularly those in the social science arena who are likely to use official statistical data at some time in the future, to have some exposure to them during their training. If this is combined with staff from the NSO being available to explain the data set and the collection objectives, methodology and editing processes, there is a great deal of scope for interaction between students and official statisticians that will be of interest and value to both parties.

The processes that a NSO applies to its data sets once the data has been collected are important to researchers, as these will have relevance to the analytical questions that researchers are investigating. Researchers need to understand that editing is undertaken to reduce discrepancies in the data but this may not result in a perfect data set. Edits are usually undertaken to meet the NSO's main aims of producing aggregated statistics and are unlikely to address the specific requirements of research, which may, for example, involve detailed modelling.

The data set could contain erroneous data, caused by keying and transposition errors, which don't impact sufficiently on the use by the NSO for them to worry about completely eradicating them. A researcher needs to know that it is possible for a data set to contain a respondent who is recorded as being born in 1937 but who might actually have been born in 1973. An inexperienced researcher is likely to be unaware of the possibility of such problems and uncertain how to deal with them when they become apparent. They will need to get advice on the potential impact of such matters on their research and then decide how material they will be to their research and determine the potential impact on their proposed outcomes.

SNZ has found it is essential for researchers to have detailed discussions with its subject matter experts to ensure they understand data set contents. Researchers normally

require assistance to determine which variables are most appropriate to their needs, and they will certainly need details on the coding schema. Being familiar with the collection instrument, often a survey questionnaire allows the researcher to determine what the variables represent. Researchers must also be aware of the collection design to ensure their proposed use of the data is supported by what is collected. This is discussed in more detail in section 5 below.

Researchers need to be willing to discuss their research and to take advice on an appropriate data set formats from subject matter experts. This requires skills in communication and the capacity to be flexible about their research plans based on advice received. While it takes time and effort to understand what is collected and the quality and limitations of the collections, researchers that SNZ have worked with have found that this process has assisted them to clarify their ideas, as well as compelling them to put clear boundaries on their project. SNZ also benefits from finding out about the researchers proposed research.

There are other benefits for SNZ. In the past documentation practices have been based solely on internal needs, with staff operating in an environment where systems and processes didn't change very quickly. Faced with increasing numbers of external researchers it was apparent we needed to provide them with well documented information about collections that was produced with a different audience in mind. With developments in technology, change happens quickly so a more rigorous approach to documentation is essential for SNZ to operate effectively. It is also of benefit to researcher, and now meta-data (questionnaires, descriptions of collections, variable lists, quality issues, contact people, and classification schemes) is available on SNZ's web site.

5. SURVEY DESIGN AND DATA QUALITY

It is not intended here to discuss training in the more technical areas of statistical analysis, such as sampling. The need for such training is not disputed and is well covered in past discussions on researchers' educational needs (e.g. Chambers & Skinner, 1998; Jolliffe, 1998; Manly & McDonald, 1998) and many efforts are being made to ensure students improve these skills.

What I will focus on is the implications of such issues within an official statistical perspective. Glencross and Mji (2001) mention the importance of validity and the need to ask "Does the instrument measure what it is supposed to measure?". A researcher using official statistical data, or any data set that already exists, needs the skills to ask, "Did the instrument measure what I needed it to measure?"

Occasionally SNZ encounters the perception that because the data has been collected by the NSO it has a very high degree of precision. In reality, there are quality limitations based around the sample design and collection instrument(s), which limit the types of analyses that can be undertaken. Inferences based on results from very small samples may not be possible because of associated large sampling errors, so while a data set appears promising it may transpire that it is not valid for the desired purpose.

As Jolliffe (2001) points out researchers need to beware of what techniques are appropriate. SNZ's surveys are not usually undertaken using simple random samples but invariably employ a complex survey design. The design will have been developed for the purpose of producing the NSO's primary outputs and therefore may not be entirely suitable for the researcher's purpose. This means that the researcher needs to understand

the relationship between their objectives and the data collection objectives and design, and needs to use analytical techniques that take the complex design into account.

Researchers also need an appreciation of non-sampling errors in general, and the issues that are specific to each data set, including both the level and composition of non-responses, and frame issues such as coverage of population and diminishing quality over time. Questionnaire design may also result in unknown biases in the data obtained, so some experience with the possible effects of the way a question is asked would be useful background for a researcher.

The issues themselves are not unique to official statistical collections and by inference can be extended to the use of many other data sets. However the way that NSO's respond to issues is determined by the legislative and political environment within which they operate. The most effective way for researchers to be exposed to these issues is to gain experience with real data sets and, in the case of official statistical data sets, to get input and assistance from NSO staff on how they have dealt with design and collection matters.

6. SIZE AND STRUCTURE OF STATISTICAL DATA SETS

As discussed earlier, one technique for reducing the risk of disclosure in a unit record data set is to provide only the subset of variables that the researcher needs to undertake their research. Additionally limiting the size of the data set is valuable on purely practical grounds. Some of SNZ's data sets are very large and complex; e.g. the population census has several million records and about 200 variables once all the derived variables are taken into account. There are often variables that superficially may appear to be the same (for example, Mäori ethnicity and Mäori ancestry) but are, in fact, quite different and valid for use in different circumstances.

Some of the data sets that SNZ has made available to researchers have been as large as one gigabyte. By the time they derive additional variables and do various sorts the data sets become unwieldy and time-consuming to handle. Jolliffe (2001) contends that problems with today's computing power computational problems are less of a problem than in the past. However our experience has been that the size of statistical data sets is still much larger than many researchers are used to and, by approaching their analyses in an inappropriate way, they can still adversely affect the performance of a large organisations computer system, as SNZ has experienced. The guidelines we develop for our own staff in efficient ways to use system resources have been provided to Data Laboratory researchers, with hints on how to sort efficiently and advice on suitable commands to use.

To be efficient a researcher should also be experienced with their analysis software. Sometimes, however use of software familiar to the researcher is not possible. Some software packages have limitations, for instance on the number of records that can be stored. As well as coming to grips with the data, a researcher may also be faced with using new software that may not have the options that they are familiar with. This adds to the demands on a researchers time, as well as their skills and knowledge. Jolliffe (2001) suggests that courses should ensure that researchers understand the principles of using software rather than the specific commands in whatever package is used for the training.

There are implications for staff in an agency who are called on to provide advice to researchers. If the researcher is using software that staff don't know they will be less

likely to be able to provide useful tips to improve the performance or advice on whether the output is successfully producing expected results.

The way data sets are stored (for example, in a hierarchical structure) or the processes used to manipulate the data sets also needs to be taken into account. These are all issues that the researcher needs to clarify at the start of the project. This means that the researcher cannot jump straight into the number crunching that is likely to be their most interesting aspect of the work. Practical experience with large data sets and the software typically used to manipulate and analyse them would be useful to researchers.

7. QUANTITATIVE SKILLS

While researchers usually have good theoretical research and analytical skills they do not always have strong programming or quantitative skills, i.e. they may know what the problem is and how to interpret the results but may not have the skills to produce the results by manipulating the data directly. In some cases they will employ a research assistant to provide programming services. However there is an opportunity for staff from the data collection agency who are likely to have strong quantitative skills, to collaborate with the researcher.

SNZ has been involved in several projects in this way. Successful collaborative research benefits both parties. A researcher learns more about the data and is more likely to acquire a clearer perspective on issues of data collection and processing. Collection agency employees are exposed to 'real world' uses of their data, which will benefit their own work. Indeed, collaborative research between the researcher with the theoretical expertise and the statistician with the data expertise may even result in a better research outcome than the researcher could achieve on their own. Svensson (1998) showed that participants in a training course on biostatistics found a multi-professional approach to be valuable.

SNZ has certainly found that its staff welcome the opportunity of working alongside experienced researchers. The researchers also appreciate the skills and knowledge that the staff member brings. This collaboration helps to strengthen the relationship between the NSO and the research sector and SNZ is actively encouraging such projects where it is appropriate.

8. OTHER SKILLS

Finally there are the skills that a researcher needs that are not specifically related to statistical research but which are an important part of successfully undertaking a research project. Some, such as communication, have been mentioned already and are addressed in other conference papers. For instance Jolliffe (2001) suggests improving researchers skills through requiring written and oral presentations as part of their courses. SNZ includes a mock interview with a 'client' as part of its in-house training on sample design for newly employed statisticians. Jolliffe also suggests that a researcher should know how to consult, i.e. ask relevant questions, and we have seen how important that is where the researcher is using a data set that already exists.

Other skills that researchers are likely to find useful are project management skills to ensure their efforts deliver results effectively. Glencross and Mji (2001) identify promotional skills as useful to ensure research results are widely disseminated, however

promotional, writing and negotiation skills also have a place at an early stage of a project if a researcher needs to secure funding to undertake their research.

9. RECOMMENDATIONS FOR TRAINING OF RESEARCHERS

The issues addressed in this paper have implications for the training and skills that social science researchers would find beneficial if they were considering undertaking research using official statistical micro data. However the conclusions are wider than just for data sets collected by NSO's and apply equally to other agencies' data sets.

It is recommended that university courses on quantitative research include a section on the use of existing data sets, in particular official statistics, and the pros and cons of micro data, e.g. using hierarchical data sets, dealing with missing values and non-response, with imputed and edited data, appropriate use of weights, analysis of complex sample design, and confidentiality control.

Practical experience with large and realistic and possibly less sanitised, statistical data sets during their training would be useful for researchers to develop an understanding of possible difficulties. Previous exposure to these issues would prevent them wasting valuable time whilst undertaking their research. It would useful for researchers to have the opportunity either to become familiar with software commonly used with large data sets or are trained in the principles of software applications rather than the specific commands of a particular package.

SNZ would support the benefits to be gained by NSO staff assisting with researcher training. It would expose NSO staff to the next generation of researchers, raise the awareness of the NSO and its data resources, and improve interaction between the research and government sectors.

There is scope for more co-operations between the sectors, for instance, in the form of the co-operative resource centres that Glencross and Mji suggest. However, researchers need to understand and be comfortable working in a co-operative research environment and to work collaboratively researchers need to be willing to recognise and understand interests other than their own.

Their training should expose them to the influences affecting government, business and the wider community and the need to balance research objectives against practical considerations of the rights of individual respondents to have their data appropriately protected and used. The resource difficulties that government agencies experience and the legislative environment that controls and constrains agencies are also relevant to what data can be made available and in what forms. SNZ has found it is particularly important that a researcher understands the particular position that it must take on some occasions, especially where that may be less favourable to the researcher.

Researchers need good communication and negotiating skills to be able to write a clear brief on their research, and to be able to defend their requirements while being willing to compromise. These are similar to skills required for writing proposals to obtain research funding.

As we have seen though the NSO can also learn from their interaction with the researcher. SNZ's practices in documentation and in data management are changing as a result of our relationship with researchers.

REFERENCES

Biggieri, L., & Zuliani, A. (1999). The dissemination of statistical literacy among citizens and public administrators. *Proceedings of 52nd Session of International Statistical Institute. Bulletin of the International Statistical Institute. Bulletin of the International Statistical Institute* (http://www.stat.fi/isi99/proceedings.html). Helsinki: International Statistical Institute.

Bishop, G., & Talbot, M. (2001). Statistical thinking for novice researchers in the biological sciences. In C. Batanero (Ed.), *Training Researchers in the Use of Statistics* (pp. 215-226). Granada: International Association for Statistical Education and International Statistical Institute.

Chambers, R., & Skinner, C. J. (1998) Communicating sampling concepts to social scientists: The CASS experience. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics*. (pp. 259-266). Singapore: International Association for Statistical Education and International Statistical Institute.

Chant, D., & Lievesley, D. (1997). The use of data in teaching statistics. *Proceedings of the 51st Session of ISI. Bulletin of the International Statistical Institute* (Tome LVII, Book 1, pp 429-432). Istambul: International Statistical Institute.

Glencross, M., & Mji, A. (2001). The role of a research resource centre in the training of social science researchers. In C. Batanero (Ed.), *Training Researchers in the Use of Statistics* (pp. 245-257). Granada: International Association for Statistical Education and International Statistical Institute.

Jolliffe, F. (1998) A course on sample surveys for statistics students. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics*. (Vol. 1, pp. 267-272). Singapore: International Association for Statistical Education and International Statistical Institute.

Jolliffe, F. (2001). Learning from experience. In C. Batanero (Ed.), *Training Researchers in the Use of Statistics* (pp. 355-370). Granada: International Association for Statistical Education and International Statistical Institute.

Manly, B. F. J., & McDonald, L. L. (1998) Teaching sampling methods using ecological examples. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics*. (Vol. 1, pp. 273-278). Singapore: International Statistical Institute.

Svensson, E. (1998) Teaching biostatistics to clinical research groups. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching of Statistics*. (Vol. 1, pp. 299-293). Singapore: International Association for Statistical Education. International Association for Statistical Education and International Statistical Institute.

Sandra McDonald
Statistics New Zealand,
Wellington, New Zealand
E-mail: sandra_mcdonald@stats.govt.nz